DOCUMENT RESUME

ED 362 524 TM 020 374

AUTHOR Keselman, Joanne C.; And Others
TITLE The Analysis of Repeated Measurements: A Quantitative Research Synthesis.
PUB DATE Apr 93
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
AVAILABLE FROM Office of Research Administration, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Computer Simulation; Foreign Countries; Interaction; Least Squares Statistics; Literature Reviews; *Mathematical Models; Meta Analysis; *Monte Carlo Methods; Multivariate Analysis; Regression (Statistics); *Research Design; Robustness (Statistics); Statistical Studies; *Synthesis
IDENTIFIERS F Test; Generalized Approximate Procedure; *Repeated Measures Design; Sphericity Tests; Type I Errors

ABSTRACT

Meta-analytic methods were used to summarize results of Monte Carlo (MC) studies investigating the robustness of various statistical procedures for testing within-subjects effects in split-plot repeated measures designs. Through a literature review, accessible MC studies were identified, and characteristics (simulation factors) and outcomes (rates of Type I error) of each MC study were coded for univariate, df-adjusted univariate, and multivariate test procedures. Results of weighted least squares regression indicate that all procedures are generally robust to violations of the multivariate normality assumption. The e-circumflex F-test was generally insensitive to departures from the sphericity assumption, with degree of bias decreasing with increases in the degree of non-sphericity. For balanced designs, all test procedures were reasonably robust to moderate degrees of covariance heterogeneity, with Type I error rates becoming only slightly liberal with increases in the degree of covariance heterogeneity. When the design was unbalanced, however, all procedures were sensitive to the presence of heterogeneous covariance matrixes, particularly for the within-subjects interaction effect, where the tests became increasingly conservative or liberal depending on the pairing of unequal covariance matrices and unequal group sizes. For balanced designs, either a df-adjusted univariate or multivariate approach is recommended. For unbalanced designs, the generalized approximate (GA) or improved GA procedures of H. Huynh (1978) are recommended rather than the investigated test procedures. Eleven tables summarize analyses. (Contains 49 references.) (SLD)

ED 362 524

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JOANNE KESELMAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE ANALYSIS OF REPEATED MEASUREMENTS: A QUANTITATIVE RESEARCH SYNTHESIS

Joanne C. Keselman
Department of Educational Psychology
University of Manitoba

Lisa M. Lix
Department of Psychology
University of Manitoba

H. J. Keselman
Department of Psychology
University of Manitoba

Paper presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia, April 12 - 16, 1993

Requests for copies of this paper should be directed to: Joanne C. Keselman, Office of Research Administration, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

KEYWORDS: Repeated measures, Split-plot, Assumption violations, Type I error, Monte carlo, Meta-analysis

TM 020374

ABSTRACT

Meta-analytic methods were used to summarize the results of Monte Carlo (MC) studies investigating the robustness of various statistical procedures for testing within-subjects effects in split-plot repeated measures designs. Through an extensive review of the literature, a population of accessible MC studies was identified, and the characteristics (simulation factors) and outcomes (rates of Type I error) of each MC study were coded for several test procedures (univariate, df-adjusted univariate, multivariate). Weighted least squares regression procedures were used to model the variation in the rates of Type I error of the test procedures as a function of the MC study characteristics. Results indicated that all test procedures were generally robust to violations of the multivariate normality assumption. The ê F-test was generally insensitive to departures from the sphericity assumption, with its degree of bias decreasing with increases in the degree of non-sphericity. For balanced designs, all test procedures were reasonably robust to moderate degrees of covariance heterogeneity, with Type I error rates becoming only slightly liberal with increases in the degree of covariance heterogeneity. When the design was unbalanced, however, all procedures were sensitive to the presence of heterogeneous covariance matrices, particularly for the within-subjects interaction effect, where the tests became increasingly conservative or liberal depending on the pairing of unequal covariance matrices and unequal group sizes. For balanced designs, the use of either a df-adjusted univariate or multivariate approach is recommended; for unbalanced designs, Huynh's (1978) GA or IGA procedures are recommended rather than any of the investigated test procedures.

THE ANALYSIS OF REPEATED MEASUREMENTS: A QUANTITATIVE RESEARCH SYNTHESIS

Box (1953) coined the term "robustness" to refer to the insensitivity of a statistical test's rates of Type I error and power to violations of its derivational assumptions. That is, "A 'robust' statistical test preserves the validity of the probability statements applied to it even though the assumptions upon which it is based are violated" (Glass, Peckham & Sanders, 1972, p. 284). According to Box (1953), the robustness of a statistical test is perhaps the most important criterion to be considered when researchers assess the usefulness of various competing statistical procedures. Thus, it is not surprising that, since Box's (1953) paper, extensive literatures have developed regarding the robustness of statistical procedures that are frequently used by educational and psychological researchers. Some of the studies comprising these literatures have used exact statistical theory to study the effects of assumption violation; however, as educational and psychological research data seldom satisfy the assumption(s) required to use exact statistical theory (i.e., normality) (Micceri, 1989), these literatures are made up, in large part, of empirical studies which have used computer simulation [or Monte Carlo (MC) methods] to examine the issue of statistical robustness. Typically, the results of these MC studies have been analyzed in a descriptive and impressionistic fashion, without the benefit of an "overarching theory" to guide their interpretation (Harwell, 1992). As a result, both methodological and applied researchers have arrived at different conclusions about the robustness of various statistical procedures on the basis of the results of a MC study or a series of MC studies. For example, on the basis of their review of the statistical robustness literature regarding the one-way fixed effects analysis of variance (ANOVA) F-test, Glass et al (1972) concluded that this test statistic was robust to violations of the homogeneity of population

variances assumption, provided that group sizes were equal. Others, however, have drawn different conclusions from this literature regarding the robustness of the ANOVA F-test (e.g., Blair, 1981) and subsequent research has questioned further the validity of Glass et al's (1972) conclusions (e.g., Rogan & Keselman, 1977; Tomarkin & Serlin, 1986).

In order to address this situation, Harwell (1992) suggested that meta-analytic methods (Glass, 1976; Glass, McGaw & Smith, 1981) be used to summarize MC research results. Within this context, the objective of a meta-analysis is to correctly model the variation in the empirical Type I/power rates of a given test procedure (the outcome variables) as a function on the characteristics of the MC studies (the explanatory variables) and, in doing so, to isolate important relationships between MC study characteristics and Type I error/power rates. For example, Harwell, Rubenstein, Hayes and Olds (1992) conducted a meta-analysis of MC studies of the robustness of the fixed effects ANOVA F-test, in which they examined variation in the rates of Type I error and power of the ANOVA F-test as a function of a number of MC study characteristics, including population shape (i.e., skewness and kurtosis), number of treatment groups, total sample size, ratio of largest to smallest group size, ratio of largest to smallest group variance and pairing of unequal group sizes and unequal group variances.[1] On the basis of the results of their meta-analysis combined with the results of exact statistical theory, Harwell at al (1992) concluded, among other things, that the ANOVA F-test is not robust to variance heterogeneity when group sizes are equal, a finding contrary to that based on the narrative review of Glass at al (1972).

---

[1] Harwell at al (1992) also investigated the Welch and Kruskal-Wallis tests in the single factor ANOVA model

According to Harwell (1992), meta-analytic procedures are particularly suitable for synthesizing MC research results. That is, because of the limited number and the nature of factors generally investigated in MC studies as well as the control that is exercised over the data generation process, MC studies do not seem to suffer from the problems that are often inherent in meta-analyses in other areas of the behavioral sciences, e.g., study selection biases, lack of comparability across studies in the definition and measurement of explanatory and outcome variables (for a detailed discussion of issues related to the use of meta-analytic methods to synthesize MC results, see Harwell, 1992, p.300-306). In addition, conducting a meta-analysis of MC research results has several advantages. Such an analysis results in the generation of a "network" of empirical results which provides a framework within which past and future MC studies can be interpreted (Harwell, 1992). This, in turn, should allow for the development of more comprehensive and valid guidelines for applied researchers than those currently available regarding the appropriate use of popular statistical procedures under specific assumption violation conditions. The integration of MC research results can also suggest avenues for further empirical work, by highlighting gaps in the extant methodological literature.

One statistical robustness literature that would profit from a quantitative research synthesis is that concerning the (omnibus) analysis of repeated measures designs. In a typical repeated measures design, subjects (or sampling units) are selected randomly for each combination of the between-subjects factors (for designs containing at least one between-subjects factor) and are exposed to each combination of the within-subjects factors (Winer, 1971). The data from a repeated measures design can be analyzed using either univariate or multivariate procedures. The valid use of either approach, however, depends on the data conforming to the

derivational assumptions underlying these procedures. Under normality, the assumptions underlying the use of a traditional mixed-model univariate approach are: (1) equality or homogeneity of covariance matrices for each suitably chosen set of orthonormalized variables at all levels of the between-subjects factors, and (2) sphericity of the common covariance matrix. Jointly, these conditions have been referred to as multisample sphericity (Huynh, 1978; Mendoza, 1980). The valid use of a multivariate approach rests on the same assumptions as that of the mixed-model univariate approach except that it does not depend on the form of the common covariance matrix, in other words, sphericity of the common covariance matrix is not required.

Unfortunately, educational and psychological research data seldom satisfy the assumptions underlying the valid use of these test procedures (Davidson, 1972; Greenwald, 1976). Consequently, and as a result of the popularity of these designs among educational and psychological researchers, an extensive literature has developed on the proper analysis strategy for such designs. In large part, this literature has focused on the effects of violating the seldom-satisfied assumption of (multisample) sphericity on the operating characteristics of the traditional mixed-model univariate F-test and on the relative robustness of various suggested alternative analysis strategies [e.g., df-adjusted univariate tests (Greenhouse & Geisser, 1959; Huynh & Feldt, 1976); multivariate tests (O'Brien & Kaiser, 1985; Timm, 1975) to this assumption violation. Some of the studies comprising this literature have used exact statistical theory to examine the robustness of these procedures. The great majority of these studies, however, have employed MC methods, with each study investigating a variety of data-analytic approaches under particular sets of simulation conditions and, as a consequence, with studies often providing

competing recommendations to applied researchers concerning appropriate analysis strategies. Indeed, Muller and Barton (1989), in summarizing the statistical robustness literature on repeated measurements, commented on the lack of consensus concerning which analysis strategy to use in any particular situation.

The purpose of the present study, therefore, was to conduct a meta-analysis of the statistical robustness literature on the analysis of repeated measures designs. Of particular interest was the integration of MC research findings concerning the use of univariate, df-adjusted univariate and multivariate procedures to analyze within-subjects effects in split-plot repeated measures designs. The results of the meta-analysis are compared and combined with those of exact statistical theory in order to arrive at a summary of the effects of assumption violations for the various test procedures. On the basis of this summary, guidelines for researchers concerning the analysis of repeated measurements under conditions of assumption violation are presented.

METHOD

Data Collection and Evaluation

A population of accessible MC studies on the use of univariate, df-adjusted univariate and/or multivariate procedures for the analysis of within-subjects effects in split-plot repeated measures designs was identified by conducting computerized and/or manual searches of the following data bases: ERIC (1969-1992); Conference Papers Index (1973 - 1992); Dissertation Abstracts International (1960-1992); PsychINFO (1967-1992); and MATHSCI (1959-1992). The following keywords were used to identify relevant studies: repeated measure, omnibus tests, tests of mean equality, nonsphericity, assumption violations, Type I error, power, Monte Carlo, and

simulation. In addition, the reference lists of all relevant studies identified through these searches were reviewed to locate articles that may have been overlooked.

The search procedures resulted in the identification of 32 studies, 17 published journal articles and 15 unpublished conference presentations, dissertations or manuscripts. Two of the identified studies did not report empirical Type I error or power rates and, therefore, were excluded from the meta-analysis. Eighteen of the remaining 30 studies reported data for repeated measures designs involving a between-subjects grouping factor. Two of these studies were eliminated from the meta-analysis as they involved a blocking variable on the within-subjects factor. One study was also eliminated from consideration as the empirical rates reported in this study represented the average of rates reported in one of the dissertations identified as part of the search procedures.

In short, the population of accessible MC studies concerning the analysis of split-plot repeated measures designs was comprised of 15 studies. All of these studies investigated analysis procedures for the simplest of split-plot repeated measures designs, that is, for designs containing a single between-subjects factor A with $j = 1,\ldots,J$ levels and n_j observations at each j ($\Sigma n_j = N$) and a single within-subjects factor B with $k = 1,\ldots,K$ levels. Eight of the studies reported only Type I error data while the remaining seven studies reported both Type I error and power data. Eight of the 15 studies were published and seven were unpublished. The seven unpublished studies consisted of three dissertations, three conference presentations, and one manuscript. As the population of accessible MC studies was relatively small, it was decided to use all 15 studies in the meta-analysis (see the appendix for a complete listing of these studies). Given the relative homogeneity of MC studies (Harwell,1992), the resulting sample of studies,

though potentially a nonrandom sample, was felt to be representative of the population of studies of interest.

Each of the 15 studies were screened for methodological flaws. In addition, the method of data generation used in each study and the accuracy of the method, as evidenced by the pattern of the test procedures' empirical rates under conditions of no assumption violations, was reviewed. On the basis of this review, all studies were judged to be methodologically sound and, therefore, all 15 studies were included in the meta-analysis.

Table 1 presents the characteristics of each study (explanatory variables) that were coded in an attempt to correctly model the variation in the investigated test procedures' empirical rates of Type I error (outcome variable)[2,3]. In order to ensure that these study characteristics were coded accurately, each MC study was reviewed and coded independently by two of the authors and any differences over the specific characteristics investigated in a given MC study were resolved jointly. The percent of agreement between the reviewers with respect to the characteristics investigated in each MC study was nearly 100%.

As seen from Table 1, information about the shape of the population distributions was captured by coding the specific type of distribution (e.g., normal, chi-square) as well as by computing values of skewness and kurtosis for these distributions (see Hastings & Peacock, 1975). To capture information about the form of the overall population covariance matrix,

---

[2] Only the results of the meta-analysis on Type I error rates ($\alpha = .05$) are presented in this paper.

[3] The variable representing the number of Monte Carlo samples, NREPS, was coded in order to compute the weights used in the weighted least squares regression analysis

Box's (1954b) correction factor, ϵ, was used where

$$\epsilon = \frac{(\mathrm{tr}C'\Sigma C)^2}{(K-1)\mathrm{tr}(C'\Sigma C)^2} \quad (1)$$

In equation (1),

$$\Sigma = \frac{\sum_{j=1}^{J} n_j \Sigma_j}{N}, \quad (2)$$

where Σ_j is the covariance matrix for group j, C is a matrix of coefficients defining a set of K-1 orthonormalized variables associated with the B and A×B within-subjects main and interaction effects, respectively, and 'tr' is the trace operator. According to Box (1954b), when the sphericity assumption is satisfied, that is, when Σ is spherical in form, $\epsilon = 1.00$; with increasing departures from sphericity, the value of ϵ decreases from 1.00 to a lower bound of $\epsilon = (K - 1)^{-1}$.

Two indices were used to capture information about the pairing of group sizes (n_j) and group covariance matrices (Σ_j). Both of these indices are based on Box's (1954a) findings concerning the effect of variance heterogeneity on the F-test in the one-way independent groups design. According to Box (1954a), when group sizes are equal, the extent of discrepancy between the empirical and nominal rates of Type I error of the F-test is a function of the spread of the distribution of unequal variances, as measured by a coefficient of variation of these variances. When group sizes are not equal, this discrepancy depends not only on this coefficient of variation but also on the ratio of the unweighted and weighted means of these variances, as reflected in Box's (1954a) bias coefficient (p.301). Applying Box's (1954a) findings to repeated measures designs involving a between-subjects grouping factor, the first index used to capture

information about the pairing of n_js and Σ_js was a weighted coefficient of variation of the group covariance matrices, WCV(GPCOV), where

$$WCV(GPCOV) = \frac{\sqrt{\sum_{j=1}^{J} n_j(\Delta_j - \bar{\Delta}_w)^2/J}}{\bar{\Delta}_w} \quad (3)$$

In equation (3), Δ_j is the difference between the average of the K variances and the average of the K(K-1)/2 covariances for group j, that is, $\Delta_j = \bar{\sigma}_{jk}^2 - \bar{\sigma}_{jkk'}$ $(k \neq k')$ and

$$\bar{\Delta}_w = \frac{\sum_{j=1}^{J} n_j \Delta_j}{N} \quad (4)$$

When group covariance matrices are homogeneous (and group sizes are either equal or unequal), WCV(GPCOV) = 0.0; as group covariance matrices become increasingly heterogeneous, the value of WCV(GPCOV) increases from zero and is a function of both the degree of covariance heterogeneity and group size inequality and of the nature of the pairing of unequal covariance matrices and unequal group sizes. For fixed degrees of covariance heterogeneity and group size inequality, the value of WCV(GPCOV) is larger for conditions in which the unequal covariance matrices are inversely paired with unequal group sizes than for conditions involving the direct pairing of these covariance matrices and group sizes[4].

The second index used to capture information about the pairing of n_js and Σ_js was a

---

[4] A direct pairing of unequal covariance matrices and group sizes refers to the case where the covariance matrix of the smallest group has the smallest Δ_j while an inverse pairing of unequal covariance matrices and group sizes refers to the case where the covariance matrix of the largest group has the smallest Δ_j.

modified version of the Box's (1954a) bias coefficient, where

$$BIAS = 1 + \frac{1 - \frac{1}{N}}{1 - \frac{1}{J}} \left[\frac{\bar{\Delta}_{uw}}{\bar{\Delta}_{w}} - 1 \right] \quad , \tag{5}$$

where

$$\bar{\Delta}_{uw} = \frac{\sum_{j=1}^{J} \Delta_j}{J} \quad , \tag{6}$$

and $\bar{\Delta}_{w}$ is as previously defined. When group covariance matrices and/or group sizes are equal, BIAS = 1.00; when both group covariance matrices and group sizes are unequal, the value of BIAS $\neq$ 1.00 and, for a fixed value of N and J, is a function of the degree of covariance heterogeneity, the degree of group size inequality and the nature of the pairing of unequal covariance matrices and unequal group sizes. When $\bar{\Delta}_{uw} < \bar{\Delta}_{w}$, as is the case for direct pairings of unequal group covariance matrices and unequal group sizes, the value of BIAS is less than 1.00 and decreases from 1.00 with increases in the degree of covariance heterogeneity and/or group size inequality. When $\bar{\Delta}_{uw} > \bar{\Delta}_{w}$, as is the case when unequal group covariance matrices are inversely paired with unequal group sizes, the value of BIAS is greater than 1.00 and increases from 1.00 with increases in the degree of covariance heterogeneity and/or group size inequality.

As seen from Table 1, information about group covariance matrices and group sizes equality/inequality was also captured separately through the GPCOV and CV(GPN) explanatory variables, respectively, where CV(GPN) is a coefficient of variation of the group sizes, that is,

$$CV(GPN) = \frac{\sqrt{\sum_{j=1}^{J} (n_j - \bar{n})^2 / J}}{\bar{n}} \quad , \qquad (7)$$

where $\bar{n} = \Sigma_j n_j / J$. These two variables were not used as explanatory variables per se but were used to stratify the MC data in order to perform various sub-analyses. Finally, various features of the design of MC study were also coded, including the number of levels of the between-subjects factor (GROUPS), number of levels of the within-subjects factor (TRIALS) and total sample size (N).

All of the MC data were coded by one of the authors, entered into a code book and subsequently transferred into a computer file. Several procedures were used to detect and correct data entry errors. First, the computer data file was checked twice for coding errors by the author who originally entered the data. Subsequent to these reviews of the data file, a systematic sample of the data was checked in terms of the accuracy of the transfer of MC data from the original studies to the computer data file. For MC studies in which equal group covariance matrices were investigated, every fourth line of data was checked for accuracy; for those studies which investigated unequal group covariance matrices, every second line of data was checked[5].

Data Analysis

Descriptive statistics were used to arrive at summary information about the qualitative and quantitative variables used in the meta-analysis. In order to characterize the relationships

---

[5] A larger percentage of the data was checked for studies which involved unequal group covariance matrices as these studies were more complex than those involving equal group covariance matrices and, correspondingly, the computation of the associated explanatory variables was more complex.

between the MC study characteristics and the rates of Type I error for the various test procedures, weighted least squares fixed effects regression models were fitted to the empirical Type I error rates of each test procedure, following methods described in Hedges and Olkin (1985, pp.168-174)[6]. The population regression models were of the form:

$$R_n=\beta_0+x_{n1}\beta_1+x_{n2}\beta_2+\ldots+x_{nP}\beta_P+\varepsilon_n, \quad (8)$$

where R_n is the observed outcome variable (i.e., rate of Type I error), x_{nP} are the coded and centered[7] explanatory variables (i.e., the coded MC study characteristics), β_0 is a population intercept, β_P is the population regression coefficient that reflects the relationship between the 'Pth' explanatory variable and the outcome variable, and ϵ_n is a population error term. In order to arrive at specific recommendations for both balanced and unbalanced split-plot designs, the MC data was subdivided into cases where the between-group sizes were either equal or unequal and separate regression analyses were performed on each of these data sets.

Hedges and Olkin's (1985, p.171) Q_R statistic was used to test for the presence of a relationship between a set of explanatory variables and the outcome variable. Tests of the difference between Q_R statistics (ΔQ_R) were used to test "competing" explanatory models

---

6 According to Harwell (1992), ordinary least squares regression analysis is likely to be inappropriate for the meta-analysis of MC studies as the empirical rates of Type I error will have different variances when the number of Monte Carlo samples varies either within or across studies. In this case, weighted least squares regression analysis, which does not rest on the assumption of homogeneity of variance, is more appropriate.

7 As a means of addressing the issue of multicollinearity, the explanatory variables in the regression models were centred by subtracting the mean value of a given explanatory variable from each score on that variable. Multicollinearity can be particularly problematic in regression models which include product terms which tend to be highly correlated with their 'component' parts (Cronbach, 1987).

(Harwell, 1992), designed to assess the role of a specific explanatory variable in accounting for variation in the outcome variable. Tests of whether a particular explanatory model adequately explained the variation in the outcome variable, that is, tests of model specification, were conducted using Hedges and Olkin's (1985, p.172) Q_E statistic. The squared multiple correlation, adjusted for the number of explanatory variables (R^2_{adj}) (see Marascuilo & Serlin, 1988, p.661) was used to quantify the predictive power of a given explanatory model and the difference in the adjusted R^2_{adj} (ΔR^2_{adj}) was used to quantify the difference in the predictive power of various explanatory models. For a detailed discussion of the Q_R and Q_E statistics, the reader is referred to Hedges and Olkin (1985, pp. 168-174); for a more detailed discussion of the use of these procedures to synthesize Monte Carlo research results, see Harwell (1992, pp. 303-306).

The SAS (SAS Institute, 1990a,b,c) statistical software program was used to perform the statistical analyses. Only the results pertaining to the following test procedures are reported in this paper: the F-test, the ê-adjusted F-test (Greenhouse & Geisser, 1959), and Hotelling's (1931) T^2 test (for the within-subjects main effect); and the F-test, the ê-adjusted F-test, and the Pillai (1955)-Bartlett (1939) trace statistic (for the within-subjects interaction effect).

RESULTS

Table 2 presents the average Type I error rates for each of the investigated test procedures by within-subjects effect and by study. A summary of the descriptive statistics computed on the qualitative and quantitative variables used in the meta-analysis for each of the within-subjects effects is presented in Tables 3 through 6. These statistics provide information on the nature of the simulation conditions investigated in the MC studies that formed the basis

of the meta-analysis. For example, as seen from Tables 5 and 6, a large percentage of the empirical distributions generated to investigate the robustness of the various test procedures were normal in form (82.6% - 93.0%). In addition, the robustness of these test procedures has typically been investigated for designs in which the number of levels of the between- and within-subjects factors are ≤ 3 (88.6% - 99.2%) and ≤ 4 (40.0% - 94.3%), respectively. In general, the ê F-test has been investigated over a broader range of designs (with respect to the number of levels of the between- and within-subjects factors and the total sample size) than either the F-test or the multivariate procedures.

A summary of the results of the explanatory models that were fitted to the empirical Type I error rates is presented in Tables 7, 8, 9 and 10[8,9]. Preliminary analyses indicated that the SKEW and KURT variables were perfectly correlated and, accordingly, the KURT variable was dropped from all regression analyses in order to eliminate this linear dependency. For each of the explanatory models, an examination of the residuals indicated no significant violations of the normality assumption. In addition, using SAS' 'RSTUDENT' residuals (see SAS, 1990b) the data was screened for observations with undue influence for each explanatory model.

As seen from Tables 7 through 10, all of the Q_R statistics were statistically significant ($\alpha = .05$), indicating that there was a significant relationship between the set of explanatory variables of each model and the Type I error rates of the various test procedures. Similarly, the

---

[8] The log-normal data was not used in the regression analyses as (a) there was a small amount of this data and (b) this data was only available for one of the within-subjects effects of interest.

[9] Missing data were treated by listwise deletion, resulting in a reduction in the number of cases used in some of the analyses.

tabled results show that, with a few exceptions, all of the Q_E statistics were also statistically significant ($\alpha = .05$), indicating that almost all of the explanatory models that were fitted to the Type I error rates were misspecified. The size and, therefore, the significance of these statistics, however, is related to the variation in the weights[10] used in the regression analyses, which was quite large for many of the explanatory models. As a result, it was felt that the R^2_{adj} statistic, an index of the explanatory power of a given model, was a more informative statistic on which to base a discussion of the various models than either Q_R or Q_E. In addition, given that the great majority of ΔQ_R statistics were also statistically significant ($\alpha = .05$), the difference in competing explanatory models is discussed with reference to the corresponding ΔR^2_{adj} values; similarly, the relative "usefulness" (Darlington, 1968) of a given explanatory variable in modelling or predicting the variation in the empirical Type I error rates of a particular test procedure is interpreted in terms of its associated squared semi-partial correlation (SP^2_p), rather than in terms of a statistical test of its associated regression coefficient ($\hat{\beta}_p$), which typically were also statistically significant.

Model Comparison #1: Effects of Bilinear Interaction Terms

To investigate the relationship between the bilinear[11] interactions among the MC study

---

[10] The weight assigned to each observation is equal to the reciprocal of the variance of the observation. For the Type I error rates, the variance is equal to $R_n(1 - R_n)/S_n$ and the reciprocal of the variance is equal to $S_n/R_n(1 - R_n)$, where S_n is the size of the sample, i.e., the number of simulations associated with R_n, the rate of Type I error.

[11] A bilinear interaction between two explanatory variables occurs when the slope of the relationship between the outcome variable and one of the explanatory variables changes as a linear function of the scores on the other explanatory variable. According to Jaccard, Turrisi and Wan (1990), the use of traditional product terms in regression analysis examine bilinear interactions between

characteristics and the empirical Type I error rates of the various test procedures, the following models were compared:

Model 1.1: TYPEI = SKEW + EPSILON + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN

Model 1.2: TYPEI = SKEW + EPSILON + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN + 21 two-way product terms[12,13,14]

As seen from Tables 7 and 9, for the equal group size cases, all of the differences in the Q_R statistics were statistically significant , with the exception of the ΔQ_R statistic associated with the F-test of the within-subjects main effect. An examination of the associated ΔR^2_{adj} values, however, suggested that the effects of the bilinear interactions among the explanatory variables on the Type I error rates of the test procedures were negligible to small, with values of ΔR^2_{adj} ranging from .003 (for the F-test of the main effect) to .078 (for the ê F-test of the main effect). As with the equal group size cases, all ΔQ_R statistics were statistically significant for the unequal group size cases, except that associated with the F-test of the within-subjects main effect. Further, with the exception of the ê F-test of the within-subjects main effect, the ΔR^2_{adj} values associated with the Model 1.1 vs. Model 1.2 comparisons were generally similar to those reported for the equal group size cases and reflected negligible to small relationships between

---

variables.

[12] For some of the analyses, linear dependencies prevented some of the product terms from entering the model (see Tables 7 through 10).

[13] The BIAS explanatory variable is bracketed ([ ]) to indicate that for the models performed on MC data involving equal group size cases, BIAS = 0 and, hence, was not included in the model.

[14] For the unequal group size cases, the GROUPS explanatory variable was not included in the models associated with the T^2 statistic as, in this case, there was no variability in this factor.

the bilinear interactions among the explanatory variables and the Type I error rates. For the ê F-test of the within-subjects main effect, these bilinear interactions had a moderate effect on its empirical rates, as indicated by an associated ΔR^2_{adj} value equal to .254. An examination of the standardized regression coefficients and associated squared semi-partial correlation coefficients for the two-way product terms of Model 1.2 suggested that the bilinear interaction of BIAS and TRIALS accounted almost exclusively for this ΔR^2_{adj} ($\hat{\beta}_{BIASxTRIALS}$ = -.6295), $SP^2_{BIASXTRIALS}$ = .205). A subsequent analysis in which the Type I error rates of the ê F-test of the within-subjects main effect were fitted to a model containing the Model 1.1. 'main effect' explanatory variables plus the BIASXTRIALS product term confirmed this observation, with the value of ΔR^2_{adj} between this model and that of Model 1.1 equalling .0586 (.6455-.5869). Thus, for the ê F-test of the within-subjects main effect, the relationship between BIAS and rates of Type I error varied to a moderate degree as a function of the number of levels of the within-subjects factor. An examination of the mean rates of Type I error indicated that when the number of levels of the within-subjects factor was small ($B \leq 4$), the mean rates of Type I error increased with increases in the value of BIAS (.0321 - .0804). When $B > 4$, however, this positive relationship between BIAS and mean rates of Type I error was not evident, with mean rates ranging from .0342 to .0307 for increasing values of BIAS.

Turning to an examination of the 'main effect' models represented by Model 1.1, the results indicated that the degree of relationship between the MC study characteristics (assumption violations and/or design features) and the Type I error rates varied as a function of the type of test procedure and the within-subjects effect of interest, with the values of R^2_{adj} ranging from .3920 to .9726. For the equal group size cases, the Type I error rates of the F-test of both

within-subjects effects were the most strongly related to the MC study characteristics (R^2_{adj} = .8438 - .8736), followed by the ê F-test (R^2_{adj} = .7334 - .8253) and the multivariate procedures (R^2_{adj} = .4251 - .5247). For the conventional F-test, this relationship was slightly stronger for the interaction tests while for the ê F-test and multivariate tests, the relationship was stronger for the within-subjects main effect.

Like the equal group size cases, the Type I error rates of the F-test for the unequal group size cases were the most strongly related to the set of Model 1.1 explanatory variables (R^2_{adj} = .8917 -.9726) and this relationship was stronger for tests of the interaction effect than for tests of the main effect. Unlike the equal group sizes cases, however, the relationship between Type I error rates and the set of Model 1.1. explanatory variables was stronger for the multivariate tests of both within-subjects effects (R^2_{adj} = .6401 - .8849) than for the ê F-test of these effects (R^2_{adj} = .3920 - .8592). Further, for both of these test procedures, the relationship between the set of Model 1.1 explanatory variables and the Type I error rates was stronger for interaction tests than for tests of the within-subjects main effect.

In summary, the Model 1.1 results indicated that the degree of relationship between the MC study characteristics and the Type I error rates of all investigated procedures for testing both main and interaction effects was moderate to very strong and that the magnitude of this relationship varied as a function of the type of test procedure and the within-subjects effect of interest. For both the equal and unequal group size cases, the strongest relationships between the set of Model 1.1 explanatory variables and rates of Type I error were associated with the F-test, with this relationship being stronger for tests of interaction effects than for tests of main tests. For the ê F-test and multivariate procedures, the strength of the relationship between the

MC study characteristics and rates of Type I error varied as a function of the type of within-subjects effect and whether the MC data were generated from balanced or unbalanced designs.

An examination of the standardized regression coefficients and the associated squared semi-partial correlation coefficients associated with the Model 1.1 regression analyses suggested that the specific MC study characteristics differed in terms of their usefulness in predicting variation in the Type I error rates of the various test procedures and that this usefulness also varied as a function of the type of within-subject effect and the nature of the MC study design (balanced or unbalanced). The remaining model comparisons were designed to tease out the usefulness of specific MC study characteristics in modelling the variation in the Type I error rates of the various test procedures.

Model Comparison #2: Effects of Population Shape (Skewness)

To examine the relationship between the shape of the population distribution, as captured by the SKEW explanatory variable, and rates of Type I error, the following models were compared:

Model 2.1: TYPEI = EPSILON + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN

Model 2.1: TYPEI = SKEW + EPSILON + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN

As all of the MC data was normal in form for the unequal group size cases, tests of the relationship of population shape and rates of Type I error were only possible for the equal group size cases. Moreover, as a large percentage of the MC data was generated from multivariate normal distributions, the tests of this relationship should be viewed as limited.

As seen from Tables 7 and 9, for the equal group size cases, all of the Model 2.1 vs. Model 2.2 ΔQ_R statistics reflecting the inclusion of the SKEW explanatory variable in the model were statistically significant, with the exception of the ΔQ_R statistic associated with the F-test of the interaction effect. An examination of the associated ΔR^2_{adj} values, however, indicated that with the exception of the multivariate test of the within-subjects main effect (T^2), the degree of relationship between the shape of the population distribution and rates of Type I error was negligible ($.000 \leq \Delta R^2_{adj} \leq .018$), indicating that the skewness of the population distribution was not useful in predicting variation in the Type I error rates of these test procedures. The observed insensitivity of the F-test to the shape of the population distribution is consistent with the large sample analytical results of Gayen (1949, 1950) and Scheffe (1959) for the independent groups ANOVA F-test; similarly, the observed robustness of the multivariate test of the within-subjects interaction effect to population shape conforms to the large sample analytical results reported by Ito (1969) and Mardia (1971) and to the empirical results of Olson (1974) regarding the Pillai-Bartlett statistic.

With respect to the multivariate test of the within-subjects effect (T^2), the results of the Model 2.1. vs. Model 2.2 comparison indicated that there was a moderate degree of relationship between the skewness of the population and its empirical rates ($\Delta R^2_{adj} = .242$). An examination of the Type I error rates for the normal and skewed data cases indicated that these rates were only slightly inflated (mean = .0668) when compared to their normal data counterparts (mean = .0558).

In summary, results indicated that all investigated procedures for testing both within-subjects main and interaction effects were generally insensitive to the shape of the population

distribution and, accordingly, that the skewness of the population distribution was generally not a useful predictor of variation in Type I error rates.

Model Comparison #3: Effects of Non-Sphericity

To examine the relationship between departures from the sphericity assumption, as captured by the EPSILON explanatory variable, and rates of Type I error, the following models were compared:

Model 3.1: TYPEI = SKEW + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN

Model 3.2: TYPEI = SKEW + EPSILON + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN

As seen from Tables 7 through 10, with the exception of the multivariate tests of both the main and interaction effects, all ΔQ_R statistics were statistically significant for the Model 3.1 vs. Model 3.2 comparisons, indicating that the rates of Type I error for both the F- and ê F-tests varied significantly as a function of the degree of departure from the sphericity assumption. As the valid use of multivariate tests does not depend on this assumption, the nonsignificant ΔQ_R statistics and corresponding negligible ΔR^2_{adj} values for the T^2 and Pillai-Bartlett statistics for both the equal and unequal group size cases are consistent with theoretical expectations.

For the equal group size cases, the variation in the Type I error rates of the F-test of both the main (ΔR^2_{adj} = .810) and interaction (ΔR^2_{adj} = .764) effects were strongly related to departures from the sphericity assumption, as captured by the EPSILON explanatory variable. As the value of EPSILON decreased from 1.00, the Type I error rates for the F-test of the main and interaction effects increased from the nominal .05 level, with mean values becoming as large

as .171 and .189, respectively. These results are consistent with the analytical results of Box (1954b) who reported that the Type I error rates of the F-test increased from the nominal level of significance as the degree of non-sphericity increased. For the ê F-test of the within-subjects main and interaction effects, the degree of relationship between Type I error rates and EPSILON was much weaker than for the F-test and generally small in size (ΔR^2_{adj} = .120 and .111, respectively). For the equal group size cases, the mean empirical Type I error rates for the ê F-test of the main and interaction effects associated with the largest investigated degree of nonsphericity were .067 and .058, respectively, suggesting that this procedure was generally insensitive to departures from the sphericity assumption. Further, the results indicated that these Type I error rates became less biased (i.e., less deviant from the nominal α) as the value of EPSILON decreased from 1.00. This finding is consistent with the results of Collier at al (1967) and Huynh and Feldt (1976), who reported that ê became a less biased estimate of ϵ as ϵ decreased from 1.00.

For the unequal group size cases, with the exception of the F-test of the within-subjects main effect, the degree of relationship between the EPSILON explanatory variable and Type I error rates was negligible ($.009 \leq \Delta R^2_{adj} \leq .047$). While the Type I error rates of the F-test of the within-subjects main effects were strongly related to the degree of non-sphericity (ΔR^2_{adj} = .630), this was not the case for the F-test of the within-subjects interaction effect (ΔR^2_{adj} = .018). A further examination of the data indicated, however, that the negligible relationship observed between the degree of non-sphericity and the Type I error rates of the F-test of the interaction effect appeared to be an artifact of the unequal group size cases. That is, as virtually all of these data (i.e., 98%) were generated when the group covariance matrices were

heterogeneous and because of the very strong degree of relationship between Type I error rates and departures from the equality of covariance matrices assumption, there was little remaining variation to be predicted by the EPSILON explanatory variable. The nature of the unequal group size cases combined with the observed moderate degree of relationship between the Type I error rates of the F-test of the within-subjects main effect and the degree of heterogeneity of the covariance matrices could also explain the reduction in the magnitude of the relationship between the empirical rates and EPSILON for the F-test of this effect, when compared to the equal group size cases. The degree of relationship between the Type I error rates of the ê F-test of both within-subjects effects was also generally weaker for the unequal group size cases and was negligible in size ($.000 \leq \Delta R^2_{adj} \leq .047$).

In summary, results indicated that for the equal group size cases the F-test of both within-subjects effects was strongly affected by departures from the sphericity assumption and, accordingly, that EPSILON was a very useful predictor of variation in the Type I error rates of this procedure. The Type I error rates of the ê F-test of both within-subjects effects were only minimally affected by the degree of non-sphericity effects, particularly for small values of ϵ. EPSILON was of little use, therefore, in predicting variation in its empirical rates. Finally, the degree of non-sphericity was not useful in predicting variation in the Type I error rates of the multivariate tests as these procedures do not depend on this assumption. For the unequal group size cases, the relationship between the rates of the Type I error of the F-test of the main and interaction effect and the degree of non-sphericity appeared to be reduced or masked, respectively, by the relationship between empirical rates and the degree of covariance heterogeneity.

Model Comparison #4: Effects of Covariance Heterogeneity

To investigate the relationship between departures from the homogeneity of covariance matrices assumption, as captured by WCV(GPCOV) (for equal group size cases) and by both WCV(GPCOV) and BIAS (for unequal group size cases), the following models were compared:

Model 4.1: TYPEI = SKEW + EPSILON + GROUPS + TRIALS + TOTALN

Model 4.2: TYPEI = SKEW + EPSILON + WCV(GPCOV) + [BIAS] + GROUPS + TRIALS + TOTALN

As seen from Tables 7 through 10, all of the ΔQ_R statistics, with the exception of that for the ê F-test of the within-subjects interaction effect for the equal group size cases, were statistically significant, indicating that with the exception noted, the Type I error rates of all test procedures varied significantly as a function of the degree of heterogeneity of the covariance matrices. An examination of the associated ΔR^2_{adj} values, however, indicated that the degree of covariance matrix heterogeneity had a negligible effect on the Type I error rates of the F- and ê F-tests of both within-subjects effects for the equal group size cases ($.000 \leq \Delta R^2_{adj} \leq .031$). These results are consistent with the analytical work of Box (1954a) and Scheffe (1959) concerning the independent groups ANOVA F-test who reported that, when group sizes were equal, this test procedure was generally robust to moderate degrees of variance heterogeneity, as was generally the case for the MC data used in the present study. For the equal group size cases, when ϵ = 1.00, the mean Type I error rates for the F-test of the main and interaction effect ranged from .052 to .055 and from .048 to .059, respectively, across the range of values of the WCV(GOCOV) explanatory variable. The corresponding ranges of mean values for the ê F-test of the main and interaction effects were .033 to .036 and .034 to .040, respectively.

The Model 4.1 vs. Model 4.2 ΔR^2_{adj} values for the multivariate tests of both the within-subjects main (ΔR^2_{adj} = .354) and interaction (ΔR^2_{adj} .= .258) effects suggested that, for the equal group size cases, the Type I error rates of these procedures were moderately affected by the degree of heterogeneity of the covariance matrices. An examination of the Type I error rates of the multivariate test procedures indicated, however, that the rates of Type I error were only slightly inflated when covariance matrices were heterogeneous, with mean empirical values ranging from .054 to .067 and from .041 to .059 over a similar range of covariance heterogeneity for the T^2 and Pillai-Bartlett statistic, respectively. These results are consistent with the theoretical results of Ito and Schull (1964) and with the empirical results of Olson (1974), who reported that the Type I error rates of the multivariate procedures are only slightly affected by moderate degree of covariance heterogeneity when group sizes are equal.

For the unequal group size cases, the ΔR^2_{adj} values indicated that the magnitude of the relationship between the empirical rates and departures from the equality of covariance matrices assumption was moderate to strong for all tests procedures and varied as a function of the type of test procedure and the within-subjects effect of interest. For all test procedures, the degree of relationship was much stronger for the test of the interaction effect ($.743 \leq \Delta R^2_{adj} \leq .949$) than for the test of the main effect ($.274 \leq \Delta R^2_{adj} \leq .554$), indicating that the effects of departures from the homogeneity of covariance matrices assumption were stronger for within-subjects interaction tests that for main effect tests. This finding is consistent with the theoretical results reported by Huynh and Feldt (1977) with respect to the mixed model ANOVA F-test, and Belli (1988) with respect to multivariate tests of repeated measures effects. For the within-subjects main effect, the strongest relationship between empirical rates and the degree of heterogeneity of the

covariance matrices was associated with the multivariate T^2 statistic; for the interaction effect, the F-test was the most affected by departures from this assumption violation.

With the exception of the T^2 statistic, the BIAS explanatory variable was much more useful in predicting variation in Type I error rates that the WCV(GPCOV) variable. The mean Type I error rates for all test procedures as a function of values of the BIAS explanatory variable are presented in Table 11. As seen from Table 11, the Type I error rates for all procedures for testing both main and interaction effects increased with increasing values of BIAS. This relationship was particularly evident for tests of the within-subjects interaction effect where empirical rates were as large as .311, .252 and .266 for the F-test, ê F-test and Pillai-Bartlett statistic, respectively. In general, when the value of BIAS was greater than 1.00, the Type I error rates of all test procedures exceeded the nominal α; when the value of BIAS was less than 1.00, these Type I error rates were generally less than the nominal value. As mentioned previously, values of BIAS less than 1.00 are associated with direct pairings of group covariance matrices and group sizes while values of BIAS greater than 1.00 are associated with inverse pairings of group covariance matrices and group sizes. According, the Table 11 results indicated that for direct pairings of group covariance matrices and group sizes, all test procedures were generally conservative while for inverse pairings of group covariance matrices and group sizes, the test procedures were generally liberal. These findings are consistent with the results reported by Box (1954a) and Horsnell (1953) for the independent groups ANOVA F-test, by Gronow (1951) and Ramsay (1980) for the independent sample t-test, and by Ito and Schull (1964) and Belli (1988) for the multivariate tests.

In summary, the results indicated that for the equal group size cases, the F- and ê F-tests

of both within-subjects effects were generally insensitive to moderate departures from the equality of covariance matrices assumption and, accordingly, that the WCV(GPCOV) explanatory variable was only minimally useful in predicting variation in the Type I error rates of these procedures over the range of covariance heterogeneity represented by the MC data. While the multivariate tests of both within-subjects effects appeared to be more sensitive to violations of this assumption for the equal group size cases, the presence of heterogeneous covariance matrices resulted in only slightly inflated Type I error rates. For the unequal group size cases, all procedures were affected by departures from the equality of covariance matrices assumption, particularly when used for testing the within-subjects interaction effect. In this case, the BIAS explanatory variable was particularly useful in predicting variation in the Type I error rates of the three test procedures, which were generally conservative when the value of BIAS was less than 1.00 and liberal when the value of BIAS was greater than 1.00.

CONCLUSIONS AND RECOMMENDATIONS

The results of the meta-analysis combined with those of exact statistical theory suggest that the F-test, ê F-test and multivariate tests are generally insensitive to departures from the multivariate normality assumption, with Type I error rates becoming only slightly inflated when the population shape was nonnormal. With respect to the sphericity assumption, the results indicate that, for balanced designs, the ê F-test of either within-subjects effect was generally insensitive to departures from this assumption, with the degree of bias in this test procedure decreasing with increases in the degree of non-sphericity. With respect to the equality of covariance matrices assumption, the effect of heterogenous covariance matrices on the Type I error rates of all of the test procedures varied as a function of the whether the design was

balanced or unbalanced and the within-subjects effect of interest. For balanced designs, the F- and ê F-tests were generally insensitive to moderate degrees of covariance heterogeneity for testing either within-subjects effect, with Type I error rates becoming only slightly inflated with increases in the degree of departure from this assumption. When the design was unbalanced, however, all procedures were sensitive to the presence of heterogeneous covariance matrices, particularly when used to test the within-subjects interaction effect, where the tests became increasingly conservative or liberal depending on the pairing of unequal covariance matrices and unequal group sizes. For unbalanced designs, a modified version of Box's (1954a) bias coefficient, BIAS, proved to be very useful in predicting variation in Type I error rates of all test procedures, particularly when applied to the assessment of the within-subjects interaction effect.

On the basis of these findings, the following guidelines are recommended to educational and psychological researchers. For balanced designs, robust tests of within-subjects effects can generally be achieved by adopting either a df-adjusted univariate (i.e., a ê F-test) or a multivariate approach. Our preference is for a multivariate approach as it depends on a less restrictive set of assumptions that a df-adjusted univariate procedure. That is, while the meta-analytic results suggest that the ê F-test is relatively robust to departures from the sphericity assumption, it is nonetheless an approximate test, unlike multivariate tests which are exact in the presence of non-sphericity provided that their own derivational assumptions are met. In this regard, the results of the present meta-analysis suggest that multivariate tests are generally robust to departures from the multivariate normality assumption.

For unbalanced designs, none of the investigated procedures can be uniformly

recommended because of their sensitivity to departures from the equality of covariance matrices assumption when group sizes are unequal, particularly when used to test the within-subjects interaction effect. For unbalanced designs, therefore, we recommend adopting Huynh's (1978) "generalized approximate" (GA) or "improved generalized approximate" (IGA) procedure. These procedures, which are extensions of the ê F- test and ẽ F-test (Huynh & Feldt, 1976), respectively, are designed for conditions of heterogeneous covariance matrices and arbitrary group sizes.

Finally, it should be noted that these conclusions and associated recommendations are restricted to the range of conditions represented by the MC data used in the present study and to the particular test procedures investigated.

APPENDIX

MC Studies Used in Meta-Analysis

1. Belli, G. (April, 1988). Type I error rates of MANOVA of repeated measures under group heterogeneity in unbalanced designs. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
2. Boik, R. J. (1975). Interactions in the analysis of variance: A procedure for interpretation and a Monte Carlo comparison of univariate and multivariate methods for repeated measures designs. Dissertation Abstracts International, 36, 2908B.
3. Collier, R. O. J., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.
4. Hearne, E. M., Clark, G. M., & Hatch, J. P. (1983). A test for serial correlation in univariate repeated-measures analysis. Biometrics, 39, 237-243.
5. Huynh, H. (1978). Some approximate tests for repeated measurements. Psychometrika, 43, 161-175.
6. Keselman, H. J., & Keselman, J. C. (1991). The analysis of repeated measurements: Univariate tests, multivariate tests, or both? Unpublished manuscript.
7. Keselman, J. C., & Keselman, H. J. (1990). Analyzing unbalanced repeated measures design. British Journal of Mathematical and Statistical Psychology, 43, 265-282.
8. Maxwell, S. E., & Arvey, R. D. (1982). Small sample profile analysis with many variables. Psychological Bulletin, 92, 778- 785.
9. Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A monte carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures designs. Multivariate Behavioral Research, 9, 165-178.
10. Noe, M. J. (April, 1976). A Monte Carlo study of several test procedures in the repeated measures designs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

11. Quintana, S. M., & Maxwell, S. E. (April, 1985). A better-than-average estimate of e. Paper presented at the annual meeting of the American Educational Research Association, Chicago, ILL.

12. Rahmann, M. (1989). Comparison of methods for the analysis of repeated measurements experiments. Dissertation Abstracts International, 51, 4913B.

13. Rasmussen, J. L. (1989). Parametric and non-parametric analysis of groups by trials design under variance-covariance inhomogeneity. British Journal of Mathematical and Statistical Psychology, 42, 91-102.

14. Rasmussen, J. L., Heumann, K. A., Heumann, M. T., & Botzum, M. (1989). Univariate and multivariate groups by trials analysis under violation of variance-covariance and normality assumptions. Multivariate Behavioral Research, 24, 93-105.

15. Rogan, J. C. (1978). A comparison of univariate and multivariate analysis strategies for repeated measures designs. Winnipeg, Manitoba: Unpublished doctoral dissertation.

REFERENCES

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. Proceedings of the Cambridge Philosophical Society, 35, 180-185.

Belli, G.B. (April, 1988). Type I error rates of MANOVA of repeated measures under group heterogeneity in unbalanced designs. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance.". Review of Educational Research, 51, 499-507.

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika, 40, 318-335.

Box, G. E. P. (1954a). Some theorems on quadratic forms applied to the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics , 25, 290-302.

Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and correlation between errors in the two-way classification. Annals of Mathematical Statistics, 25, 484-498.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Collier, R. O. J., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently

proposed. Psychological Bulletin, 102(3), 414-417.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated measures experiments. Psychological Bulletin, 77, 446-452.

Gayen, A. K. (1949). The distribution of 'Student' t in the random samples of any size drawn from non-normal universes. Biometrika, 36, 353-369.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Biometrika, 37, 236-255.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24, 95-112.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? Psychological Bulletin, 83, 314-320.

Gronow, D. G. C. (1951). Test for the significance of the difference between means in two normal populations having unequal variances. Biometrika, 38, 252-256.

Harwell, M. C. (1992). Summarizing monte carlo results in methodological research. Journal of Educational Statistics, 17(4), 297-313.

Harwell, M. C., Rubenstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17(4), 315-339.

Hastings, N. A. J., & Peacock, J. B. (1975). Statistical distributions. London: Butterworths.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Harcourt, Brace, Jovanovich.

Horsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 40, 128-136.

Hotelling, H. (1931). The generalization of Student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Hotelling, H. (1951). A generalized t test and measure of multivariate dispersion. J. Neyman, Proceedings of the Second Berkeley Symposium on Mathematical Statistics,Berkeley, CA: University of California Press.

Huynh, H. (1978). Some approximate tests for repeated measurements. Psychometrika, 43, 161-175.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

Huynh, H., & Feldt, L.S. (August, 1977). Performance of the traditional F tests in split-plot designs under covariance heterogeneity. Paper presented at the annual meeting of the

American Statistical Association, Chicago, ILL.

Ito, K. (1969). On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. In P.R. Krishnaiah (Ed.), Multivariate analysis (Vol.2). New York: Academic Press.

Ito, K., & Schull, W. J. (1964). On the robustness of the T^2 test in multivariate analysis when variance-covariance matrices are not equal. Biometrika, 51, 71-82.

Jaccard, J., Turrisi, R., & Choi, K. W. (1990). Interaction effects in multiple regression. Newbury Park, CA: Sage.

Marascuilo, L. A., & Serlin, R. C. (1988). Statistical methods for the social and behavioral sciences. New York: Freeman.

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrika, 58, 105-121.

Mendoza, J. L. (1980). A significance test for multisample sphericity. Psychometrika, 45, 495-498.

Micceri, T. (1989). The unicorn, the normal distribution, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Muller, K. E., & Barton, C. N. (1989). Approximate power for repeated measures ANOVA lacking sphericity. Journal of the American Statistical Association , 84, 549-555.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. Psychological Bulletin, 97(2), 316-333.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 69, 894-903.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. Annals of Mathematical Statistics, 26, 117-121.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of student's t test with unequal variances. Journal of Educational Statistics, 5, 337-349.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Journal, 14(4), 493-498.

SAS Institute, Inc. (1990a). SAS/STAT user's guide (vol.1) version 6 (4th Ed.). Cary, North Carolina: SAS Institute, Inc.

SAS Institute, Inc. (1990b). SAS/STAT user's guide (vol. 2) version 6 (4th Ed.). Cary, North Carolina: SAS Institute, Inc.

SAS Institute Inc. (1990c). SAS/STAT user's guide (vol.3) version 6 (4th Ed.). Cary, North Carolina: SAS Institute, Inc.

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

Timm, N. E. (1975). Multivariate analysis with applications in education and psychology. Monterey, CA: Brooks-Cole.

Tomarkin, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

Winer, B. J. (1971). Statistical principles of experimental designs (2nd ed.). New York: McGraw-Hill.

TABLE 1

Coded Monte Carlo Study Characteristics: Type I Data (α = .05)

Explanatory Variables:

(1) Distribution (γ_1 = skewness, γ_2 = kurtosis)

1 = Normal ($\gamma_1 = 0$, $\gamma_2 = 0$)
2 = Chi-Square (γ_1 ,γ_2 depend on the parameters used)
3 = Log-Normal (γ_1 ,γ_2 depend on the parameters used)

(2) Skewness (γ_1) (SKEW)

(3) Kurtosis (γ_2) (KURT)

(4) Sphericity (ϵ) (EPSILON)

(5) Pairing of Group Sizes and Covariance Matrices(Σ_j s)

a. Weighted Coefficient of Variation of the Σ_js [WCV(GPCOV)]
b. Modified Bias Coefficient (BIAS)

(6) Inequality of Group Covariance Matrices (Σ_js)(GPCOV)

1 = equal Σ_js
2 = at least one Σ_j not equal to the other Σ_js

(7) Inequality of Group Sizes [CV(GPN)]

(8) Number of Levels of Between-Subjects Factor (GROUPS)

(9) Number of Levels of Within-Subjects Factor (TRIALS)

(10) Total Sample Size (TOTALN)

Outcome Variable:

(11) Empirical Type I Error Rate (TYPEI)

Other:

(12) Number of Monte Carlo Samples for Type I data (NREPS)

TABLE 2

Mean Type I Error Values (α = .05) for Investigated Test Procedures by Type of Within-Subjects Effect (B, AxB) and Study

Study	B Main Effect	AxB Interaction
1	*	.100 (24)
2+	*	.064 (24)
3	.063 (45)	.070 (45)
4	*	*
5	.034 (6)	.034 (6)
6	.046 (168)	.066 (168)
7	.057 (96)	*
8	.027 (15)	.027 (15)
9	.060 (18)	.059 (12)
10	.069 (192)	.118 (96)
11	.038 (36)	.040 (36)
12+	.064 (31)	.065 (31)
13	*	.055 (40)
14	*	.033 (72)
15+	.058 (90)	.057 (90)

Note: Investigated test procedures were F Test, ê F Test, T^2 (for the B main effect) and F Test, ê F Test, Pillai-Bartlett statistics (for the AxB interaction effect);
* = did not examine Type I error and/or power for investigated test procedures;
+ = Ph.D. dissertation; and
values in parentheses refer to the number of cases (n).

TABLE 3

Descriptive Statistics for Quantitative Variables:
Within-Subjects Main Effect

Variable	Mean	SD	Min	Max
F-test (n = 175)				
TYPEI	.076	.020	.041	.171
SKEW	.168	.497	0	1.63
KURT	.411	1.219	0	4
EPSILON	.694	.197	.168	1
WCV(GPCOV)	.267	.241	0	1.106
BIAS	1.043	.215	.661	1.659
TOTALN	23.246	11.659	15	47
CV(GPN)	.160	.236	0	.653
NREPS	1737.143	454.281	1000	3000
ê F-test (n = 265)				
TYPEI	.043	.021	.002	.131
SKEW	.111	.411	0	1.63
KURT	.272	1.008	0	4
EPSILON	.713	.233	.168	1
WCV(GPCOV)	.228	.247	0	1.106
BIAS	1.013	.149	.665	1.618
TOTALN	33.374	18.409	9	120
CV(GPN)	.121	.181	0	.848
NREPS	2494.340	1311.721	1000	5000
T^2 (n = 257)				
TYPEI	.058	.010	.040	.099
SKEW	.114	.417	0	1.63
KURT	.280	1.023	0	4
EPSILON	.708	.215	.400	1
WCV(GPCOV)	.344	.220	0	.649
BIAS	1.043	.229	.661	1.659
TOTALN	28.767	12.105	15	45
CV(GPN)	.213	.227	0	.653
NREPS	2529.183	1202.262	1000	5000

Note: see Table 1 for variable definition.

TABLE 4

Descriptive Statistics for Quantitative Variables:
Within-Subjects Interaction Effect

Variable	Mean	SD	Min	Max
F-test (n = 219)				
TYPEI	.091	.067	.005	.320
SKEW	.473	1.448	0	6.18
KURT	6.408	25.250	0	110.94
EPSILON	.711	.201	.168	1
WCV(GPCOV)	.213	.240	0	1.106
BIAS	1.034	.193	.661	1.659
TOTALN	26.110	16.921	10	80
CV(GPN)	.127	.220	0	.653
NREPS	1890.411	508.505	1000	3000
ê F-test (n = 285)				
TYPEI	.048	.039	.001	.346
SKEW	.363	1.284	0	6.18
KURT	4.924	22.287	0	110.94
EPSILON	.739	.229	.168	1
WCV(GPCOV)	.136	.212	0	1.106
BIAS	1.009	.124	.665	1.618
TOTALN	32.779	20.653	9	120
CV(GPN)	.085	.172	0	.848
NREPS	1947.368	653.298	1000	5000
Pillai-Bartlett (n = 155)				
TYPEI	.067	.069	.001	.369
SKEW	.636	1.682	0	6.18
KURT	8.976	29.656	0	110.94
EPSILON	.783	.234	.400	1
WCV(GPCOV)	.271	.264	0	1.155
BIAS	1.030	.247	.438	2.311
TOTALN	39.684	13.903	10	80
CV(GPN)	.171	.213	0	.544
NREPS	2045.161	263.393	1000	2500

Note: see Table 1 for variable definition.

TABLE 5

Descriptive Statistics for Qualitative Variables:
Within-Subjects Main Effect

	F-test		ê F-test		T^2	
Variable	**f**	**%**	**f**	**%**	**f**	**%**
Distribution	175	100.	265	100	257	100
Normal	157	89.7	247	93.2	239	93
Chi-Square	18	10.3	18	6.8	18	7
Groups	175	100	265	100	257	100
2	2	1.1	29	10.9	0	0
3	167	95.4	206	77.7	255	99.2
4	6	3.4	6	2.3	2	0.8
6	0	0	24	9.1	0	0
Trials	175	100	265	100	257	100
3	2	1.1	2	0.8	1	0.4
4	163	93.2	136	51.3	216	84
5	3	1.7	27	10.2	2	0.8
6	3	1.7	3	1.1	1	0.4
7	2	1.1	2	0.8	1	0.4
8	1	0.6	61	23	36	14
9	1	0.6	1	0.4	0	0
13	0	0	33	12.5	0	0
GPCOV	175	100	265	100	257	100
equal	61	34.9	115	43.4	51	19.8
unequal	114	65.1	150	56.6	206	80.2

Note: see Table 1 for variable definition. f = frequency.

TABLE 6

Descriptive Statistics for Qualitative Variables:
Within-Subjects Interaction Effect

	F-test		ê F-test		Pillai-Bartlett	
Variable	**f**	**%**	**f**	**%**	**f**	**%**
Distribution	219	100	285	100	155	100
Normal	189	86.3	255	89.5	128	82.6
Chi-Square	18	8.2	18	6.3	15	9.7
Log-normal	12	5.5	12	4.2	12	7.7
Groups	219	100	285	100	155	100
2	46	21	73	25.6	48	31
3	167	76.3	182	63.9	105	67.7
4	6	2.7	6	2.1	2	1.3
6	0	0	24	8.4	0	0
Trials	219	100	285	100	155	100
3	2	0.9	2	0.7	1	0.6
4	163	74.4	112	39.3	66	42.6
5	47	21.5	71	24.9	50	32.3
6	3	1.4	3	1.1	1	0.6
7	2	0.9	2	0.7	1	0.6
8	1	0.5	61	21.4	36	23.2
9	1	0.5	1	0.4	0	0
13	0	0	33	11.6	0	0
GPCOV	219	100	285	100	155	100
equal	105	47.9	175	61.4	49	31.6
unequal	114	52.1	110	38.6	106	68.4

Note: see Table 1 for variable definition. f = frequency.

TABLE 7

Results of Explanatory Models for the Within-Subjects Main Effect: Equal Group Sizes

Model Comparison	Test	df_R	df_E	Q_R	Q_E	R^2_{adj}	ΔR^2_{adj}
Bilinear Interactions							
1.1	F	6	103	785.34	136.04	.8438	.003†
1.2		18	91	803.29	118.10	.8465	
1.1	ê	6	158	3907.52	790.50	.8253	.078
1.2		19	145	4295.91	402.12	.9032	
1.1	T^2	6	108	162.60	133.20†	.5247	.076
1.2		16	98	194.25	101.55	.6007	
Population Shape (Skewness)							
2.1	F	5	104	781.20	140.18	.8405	.003
2.2		6	103	785.34	136.04	.8438	
2.1	ê	5	159	3898.50	799.52	.8245	.001
2.2		6	158	3907.52	790.50	.8253	
2.1	T^2	5	109	93.09	202.70	.2833	.242
2.2		6	108	162.60	133.20	.5247	
Non-Sphericity							
3.1	F	5	104	72.41	848.97	.0343	.810
3.2		6	103	785.34	136.04	.8438	
3.1	ê	5	159	3353.86	1344.17	.7049	.120
3.2		6	158	3907.52	790.50	.8253	
3.1	T^2	5	109	161.97	133.83†	.5268	.002†
3.2		6	108	162.60	133.20	.5247	
Covariance Heterogeneity							
4.1	F	5	104	770.03	151.35	.8278	.016
4.2		6	103	785.34	136.04	.8438	
4.1	ê	5	159	3892.93	805.10	.8232	.002
4.2		6	158	3907.52	790.50	.8253	
4.1	T^2	5	109	61.40	234.40	.1712	.354
4.2		6	108	162.60	133.20	.5247	

Note: † indicates nonsignificant statistics ($\alpha = .05$). Marked (†) ΔR^2_{adj} value indicates corresponding ΔQ_R statistic was nonsignificant.

TABLE 8

Results of Explanatory Models for the Within-Subjects Main Effect: Unequal Group Sizes

Model Comparison	Test	df_R	df_E	Q_R	Q_E	R^2_{adj}	ΔR^2_{adj}
Bilinear Interactions							
1.1	F	6	58	622.27	67.71†	.8917	.004†
1.2		11	53	630.19	59.80†	.8953	
1.1	ê	6	93	2986.75	3977.21	.3920	.254
1.2		17	82	4918.89	2045.06	.6455	
1.1	T^2	5	136	363.96	193.49	.6401	.094
1.2		15	126	424.77	132.68†	.7336	
Non-Sphericity							
3.1	F	5	59	220.61	469.37	.2621	.630
3.2		6	58	622.27	67.71	.8917	
3.1	ê	5	94	2635.66	4328.30	.3454	.047
3.2		6	93	2986.75	3977.21	.3920	
3.1	T^2	4	137	363.36	194.09	.6417	.002†
3.2		5	136	363.96	193.49	.6401	
Covariance Heterogeneity							
4.1	F	4	60	442.43	247.55	.6173	.274
4.2		6	58	622.27	67.71	.8917	
4.1	ê	4	95	994.39	5969.57	.1067	.285
4.2		6	93	2986.75	3977.21	.3920	
4.1	T^2	3	138	59.08	498.37	.0865	.554
4.2		5	136	363.96	193.49	.6401	

Note: see Table 7 note.

TABLE 9

Results of Explanatory Models for the Within-Subjects Interaction Effect: Equal Group Sizes

Model Comparison	Test	df_R	df_E	Q_R	Q_E	R^2_{adj}	ΔR^2_{adj}
Bilinear Interactions							
1.1	F	6	135	1490.26	205.13	.8736	.026
1.2		19	122	1548.61	146.77†	.8999	
1.1	ê	6	198	2907.27	1015.04	.7334	.050
1.2		19	185	3150.94	771.37	.7831	
1.1	PB	6	68	95.19	106.59	.4251	.038
1.2		18	56	119.77	82.01	.4629	
Population Shape (Skewness)							
2.1	F	5	136	1487.54	207.84	.8729	.001†
2.2		6	135	1490.26	205.13	.8736	
2.1	ê	5	199	2902.67	1019.64	.7335	.000
2.2		6	198	2907.27	1015.04	.7334	
2.1	PB	5	69	90.21	111.57	.4070	.018
2.2		6	68	95.19	106.59	.4251	
Non-Sphericity							
3.1	F	5	136	239.73	1455.66	.1098	.764
3.2		6	135	1490.26	205.13	.8736	
3.1	ê	5	199	2478.48	1443.84	.6226	.111
3.2		6	198	2907.27	1015.04	.7334	
3.1	PB	5	69	95.06	106.73	.4328	.000†
3.2		6	68	95.19	106.59	.4251	
Covariance Heterogeneity							
4.1	F	5	136	1438.10	257.29	.8427	.031
4.2		6	135	1490.26	205.13	.8736	
4.1	ê	5	199	2905.79	1016.52	.7343	.000†
4.2		6	198	2907.27	1015.04	.7334	
4.1	PB	5	69	45.06	156.73	.1670	.258
4.2		6	68	95.19	106.59	.4251	

Note: see Table 7 note.

TABLE 10

Results of Explanatory Models for the Within-Subjects Interaction Effect: Unequal Group Sizes

Model Comparison	Test	df_R	df_E	Q_R	Q_E	R^2_{adj}	ΔR^2_{adj}
Bilinear Interactions							
1.1	F	6	58	14292.55	363.77	.9726	.010
1.2		11	53	14443.07	213.24	.9824	
1.1	ê	6	61	7208.70	1059.82	.8592	.034
1.2		17	50	7609.77	658.74	.8932	
1.1	PB	6	61	11945.72	1398.18	.8849	.031
1.2		17	50	12509.37	834.53	.9162	
Non-Sphericity							
3.1	F	5	59	14049.91	606.40	.9551	.018
3.2		6	58	14292.55	363.77	.9726	
3.1	ê	5	62	7122.04	1146.48	.8502	.009
3.2		6	61	7208.70	1059.82	.8592	
3.1	PB	5	62	11945.58	1398.32	.8868	.000†
3.2		6	61	11945.72	1398.18	.8849	
Covariance Heterogeneity							
4.1	F	4	60	1240.94	13415.37	.0236	.949
4.2		6	58	14292.55	363.77	.9726	
4.1	ê	4	63	1399.42	6869.10	.1165	.743
4.2		6	61	7208.70	1059.82	.8592	
4.1	PB	4	63	1658.86	11685.04	.0687	.816
4.2		6	61	11945.72	1398.18	.8849	

Note: see Table 7 note.

TABLE 11

Mean Type I Error Rates as a Function of the BIAS Explanatory Variable

Effect	Test	Values of BIAS (B)					
		$B \leq .60$	$.60 < B \leq .80$	$.80 < B \leq 1.00$	$1.00 < B \leq 1.20$	$1.20 < B \leq 1.40$	$B > 1.40$
B	F	0(0)	.065(12)	.070(16)	.102(1)	.084(24)	.095(12)
	ê	0(0)	.027(14)	.036(39)	.056(18)	.065(23)	.039(6)
	T^2	0(0)	.052(26)	.052(40)	.059(14)	.065(44)	.070(18)
AxB	F	0(0)	.018(12)	.049(16)	.126(1)	.160(24)	.311(12)
	ê	0(0)	.009(6)	.023(31)	.064(10)	.106(15)	.252(6)
	PB	.026(4)	.013(9)	.024(22)	.082(6)	.129(15)	.266(12)

Note: PB = Pillai-Bartlett. Values in parentheses refer to the number of cases.